Cell, Vol. 40, 477–480, March 1985, Copyright © 1985 by MIT

Molecular Organization of the AIDS Retrovirus

Minireview

A. B. Rabson and M. A. Martin
National Institute of Allergy
and Infectious Diseases
National Institutes of Health
Bethesda, Maryland 20205

Since their original clinical description in 1981, the Acquired Immunodeficiency Syndrome (AIDS) and the AIDS-Related Complex (ARC) have become extremely important public health problems throughout the world. By the end of 1984, approximately 8,000 cases of AIDS had been diagnosed in the U.S.A. with an overall mortality of over 40%; more than 100 new cases are reported each week. A major advance in the understanding of the pathogenesis of this disease has been the isolation of a novel retrovirus from AIDS and ARC patients. Serologic studies have demonstrated the association of viral infection and clinical disease. During its propagation in vitro, the AIDS retrovirus (AIDS RV) preferentially infects and kills human T lymphocytes of the OKT4/Leu-3 subset, the same cells apparently destroyed in patients with the disease. Both seroepidemiological and in vitro data strongly suggest that the retrovirus isolated from patients is in fact the causative agent of AIDS. This virus has been variously called lymphadenopathy virus (LAV), Human T Lymphotropic Virus III (HTLV-III) and AIDS-Associated Retrovirus (ARV). We shall refer to it as the AIDS RV. With the recent publication of the complete nucleotide sequences of molecular clones corresponding to the three isolates, it is now possible to identify different viral genes and make assessments regarding their expression and structural variability.

The DNA sequence of the AIDS RV has been derived from molecular clones of both unintegrated and integrated proviruses. Wain-Hobson et al. (Cell *40*, 9–17, 1985) sequenced an integrated LAV provirus cloned at HindIII sites present in the R region of both long terminal repeats (LTRs). Ratner et al. (Nature *313*, 277–284, 1985) derived the sequences of 8.9 kb of one cloned unintegrated HTLV-III provirus, as well as portions of two additional unintegrated proviruses and one integrated provirus. Sanchez-Pescador et al. (Science 227, 484–492, 1985) sequenced various regions of four integrated ARV-2 clones and one unintegrated ARV-2 circular provirus.

The sequences of LAV, HTLV-III, and ARV-2 show general agreement in the size and organization of the AIDS RV genome. The AIDS RV is the longest retrovirus sequenced to date. The DNA provirus with two LTRs is 9734–9749 bp in length, the larger size being observed in a clone of HTLV-III containing a 15 bp duplication in the *env* gene. As shown in Fig. 1, the AIDS RV contains many characteristic features of replication competent retroviruses: LTRs, group specific antigen genes (*gag*), a gene region (*pol*) encoding reverse transcriptase as well as putative endopeptidase and integrase enzymes, and a gene encoding the virus envelope glycoprotein (*env*). Although all replication competent viruses contain these genes, each retroviral family exhibits unique structural features and its own program of gene expression. The AIDS RV is no exception. In addition to having an unusually long overlap of the *gag* and *pol* genes and no overlap of *pol* and *env* (in contrast to other retroviruses), the AIDS RV contains two novel open reading frames (Fig. 1, segments A and B) that may play a role in its unusual cytopathogenicity.

Unique Features of the AIDS Proviral DNA

LTR: The 634 bp AIDS RV LTR is similar in size to other type C mammalian retroviral LTRs. Unlike HTLV-I, which has an LTR containing R and U5 regions of 228 and 176 bp, respectively, the sizes of R and U5 of the AIDS RV LTR are 97–98 and 83–84 bp, respectively. One of the major surprises of the AIDS provirus sequence is the presence of a tRNA primer binding site (pbs) complementary (18/18 match) to $tRNA^{lys}$. All other infectious mammalian retroviruses have a $tRNA^{pro}$ pbs, except mouse mammary tumor virus (MMTV), which also contains the $tRNA^{lys}$ pbs. As is true of other retroviruses except HTLV-I, HTLV-II and BLV, the AIDS RV LTR contains a polyadenylation signal (AATAAA) within R, 19 bp 5′ to the R-U5 boundary, and a TATA sequence 22–27 bp 5′ to the presumed mRNA start site. No CAAT sequence, a feature common to several LTRs, could be identified in its usual position 60–80 bp 5′ to the mRNA cap site.

gag Gene: Retroviral *gag* proteins are synthesized in the form of a polyprotein precursor which is proteolytically processed to individual *gag* proteins. The AIDS RV *gag* region (Fig. 1) is approximately 1500 bp long and could therefore encode a polyprotein of about 500 amino acids. This size is consistent with the 53–55 kd protein detected immunochemically in AIDS RV infected cells using anti-

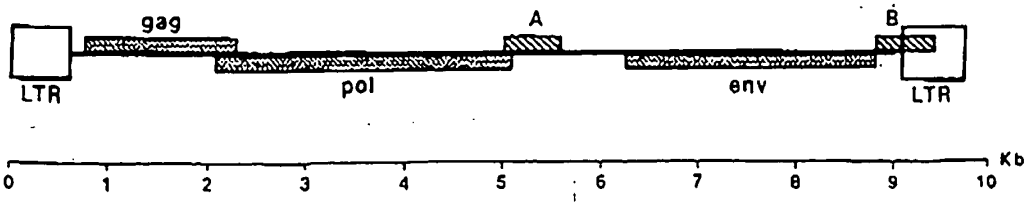

Figure 1. Organization of the AIDS Retrovirus

The gene sizes and arrangement are derived from the three published DNA sequences of the AIDS RV (see text).

body present in patient sera. In addition, p24/p25 *gag* and p17 *gag* proteins have been identified by immunoblotting or immunoprecipitation procedures. All three groups that sequenced the viral genome determined the partial amino acid sequence of one or more putative *gag* proteins; the latter were then aligned with the deduced amino acid sequence of the AIDS RV *gag* gene. Unlike the *gag* regions of infectious murine, feline, and simian retroviruses—all of which encode four *gag* proteins—the AIDS RV *gag* gene specifies only three proteins. HTLV-I and HTLV-II also have truncated *gag* genes that encode three proteins. Based on amino acid composition and amino acid sequence alignment with the *gag* regions of other sequenced retroviruses, it was concluded that the analogue of the second (of the four) *gag* proteins was absent from the AIDS RV genome.

pol Gene: The *pol* region of the AIDS RV is in a different reading frame from *gag*. It overlaps the 3' end of the *gag* gene by 80 amino acids, and is otherwise similar in organization to *pol* genes present in other mammalian retroviruses. Alignment of AIDS RV *pol* nucleotide and deduced amino acid sequences with corresponding 5' endopeptidase, reverse transcriptase, and the 3' endonuclease/integrase domains of other viruses is readily accomplished.

env Gene: The *env* region of retroviruses generally encodes a single polyprotein precursor which is cleaved to generate a larger external viral envelope glycoprotein (encoded by sequences located in the 5' portion of the *env* gene) that is attached by disulfide bonds to a smaller transmembrane protein (encoded by 3' *env* sequences). The AIDS RV possesses a very large *env* region capable of encoding an 863 amino acid precursor with an estimated unglycosylated molecular weight of 90–100 kd. Another unusual feature of the AIDS RV *env* gene sequence is the large number (30–32) of potential glycosylation sites, the majority of which (24–26) are situated upstream from the precursor cleavage site and would, therefore, be located in the external *env* protein. As a consequence, the molecular weights of the glycosylated *env* precursor and external *env* proteins could both be considerably larger than the peptides deduced from the nucleotide sequence. In fact, immunoprecipitation and Western blot analysis of AIDS RV infected cell lysates have revealed the existence of 160 and 120 kd glycoproteins.

Potential processing sites for the cleavage of the *env* precursor glycoprotein would generate an unusually large transmembrane protein containing approximately 350 amino acids. The three groups disagree about the possible function of this processed *env* protein. Ratner et al. refer to the entire AIDS RV *env* gene as the "*env-lor*" region and argue that the 3' *env* region encodes a *trans*-acting analogue of the pX-IV segment of HTLV-I. Evidence supporting the existence of *trans*-activation of LTR mediated transcription in HTLV-III infected cells seems to be quite convincing (Sodroski et al., Science 227, 171–173, 1985). However, the lack of any polynucleotide sequence homology between the *lor* segment of HTLV-I and HTLV-III proviral DNA precludes the mapping of its functional equivalent at the present time. Ratner et al. use hydrophilicity/hydrophobicity calculations to argue that the approximately 350 amino acid *env* transmembrane protein of HTLV-III is the functional equivalent of the separate 178 amino acid transmembrane and the 357 amino acid *lor* proteins encoded by the HTLV-I genome. Wain-Hobson et al. and Sanchez-Pescador et al. suggest no unusual role for the large AIDS RV *env* transmembrane protein.

An intriguing and potentially important difference between the AIDS RV and other retroviruses is the presence of two additional reading frames denoted "A" and "B" in Fig. 1. "A" is an open reading frame 196–203 amino acids long, situated between *pol* and *env*. A stretch of 580 bp of noncoding sequence begins at the 3' end of "A." In previously published sequences of other replication competent retroviruses, the *pol* and *env* reading frames overlap without intervening open reading frames or noncoding regions. It is possible that the noncoding sequences in this region of the AIDS RV have regulatory functions. A second open reading frame ("B" in Fig. 1) unique to the AIDS RV overlaps the 3' end of *env* and extends for approximately 200 amino acids, terminating about 330 nucleotides into the U3 region of the 3' LTR. This reading frame is open in LAV, ARV-2, and some clones of HTLV-III; however, it appears to contain a termination codon in at least one clone of HTLV-III. The "B" open reading frame of LAV and ARV-2 is similar to the pX-IV open reading frame of HTLV-I, which extends 78 nucleotides into the U3 region of the 3' LTR (Seiki et al., PNAS 80, 3618–3622, 1983). Furthermore, a 1.7 kb polyadenylated mRNA containing LTR and "B" region sequences has been identified in AIDS RV infected cells (Rabson et al., in preparation).

Classification of the AIDS Retrovirus

Retroviruses have been traditionally classified on the basis of their biology, electron microscopic morphology, and genomic structure.

Biological classifications have divided retroviruses into three groups: 1) the oncoviruses, many of whose members are naturally oncogenic, producing leukemias, lymphomas, and breast carcinomas; 2) the spuma viruses or foamy viruses, which produce vacuolization of tissue culture cells but no known disease; and 3) the lentiviruses, or slow viruses, which produce cytopathic effect in tissue culture cells and slowly progressive disease in animals. On biological grounds, the AIDS RV, with its capacity to produce dramatic cytopathic effect in tissue culture and slowly progressive disease in man, seems to have many features of a lentivirus. Although it is not associated with leukemia, the AIDS RV shares its target cell tropism ($OKT4^+$ human lymphocytes) and ability to form syncytia in these cells with HTLV-I and -II.

On the basis of electron microscopic morphology, the AIDS RV particle, with its bar-shaped central structure, most closely resembles visna and equine infectious anemia virus, both members of the lentivirus family.

Nucleic acid hybridization and, more recently, nucleotide sequence analysis have been useful for classifying microorganisms and establishing evolutionary relationships between closely related agents. Analysis of HTLV-III, ARV-2, and LAV nucleotide sequences clearly establishes that they are all retroviruses. Although the deduced amino

Table 1. Sequence Comparison of AIDS Retroviral Isolates

	Clones or Isolates Compared	LTR U3	LTR R	LTR U5	gag	pol	A	env	B
Nucleotide homology	HTLV-III BH10 × HTLV-III BH8	8/456 (1.8)	NA	NA	ND	ND	ND	36/2607 (1.4)	14/651 (2.2)
	HTLV-III BH10 × LAV	11/456 (2.4)	1/97 (1.0)	0/85 (0)	ND	ND	ND	46/2607 (1.8)	14/651 (2.2)
	LAV × ARV-2	30/456 (6.6)	1/97 (1.0)	1/85 (1.1)	46/1503 (3.1)	87/3012 (2.9)	32/610 (5.2)	242/2607 (9.3) 190/1300[a] (14.6)	51/651 (7.8)
Amino acid homology	LAV × ARV-2	–	–	–	17/501 (3.4)	32/1004 (3.2)	19/203 (9.4)	131/869 (15.1) 92/435[b] (21.1)	30/220 (13.5)

Values are given as nucleotide differences/nucleotides compared or amino acid differences/amino acids compared. The percentage of heterogeneity is included within the parentheses.
[a] Comparison of the 5' 1300 nucleotides of *env*.
[b] Comparison of the 5' 435 amino acids of *env*.
NA: not applicable. ND: not done.

acid sequences of the AIDS RV can be aligned with short regions of analogous coding segments present in other retroviruses, there are virtually no long stretches of polynucleotide sequence homology with other proviral DNAs. Thus, the AIDS RV genome is no more closely related to other mammalian retroviruses than it is to Rous sarcoma virus. The reported hybridization of HTLV-III with HTLV-I under low stringency conditions (Hahn et al., Nature *312*, 166–169, 1984) is difficult to explain in view of the three published nucleotide sequences of the AIDS RV. Specific duplex structures formed during the reaction of denatured HTLV-I or HTLV-II DNAs with AIDS RV DNA would be thermodynamically unstable even under low stringency hybridization conditions. In this regard, HTLV-II, which has considerable nucleotide identity to HTLV-I, does not react with LAV proviral DNA (Alizon et al., Nature *312*, 757–760, 1984).

The question, therefore, remains how the AIDS RV should be classified. A recent report by Gonda et al. (Science *227*, 173–177, 1985) indicated that HTLV-III sequences located between 0.8 and 4.6 kb (see map in Fig. 1) hybridized to a ^{32}P-labeled nick-translated visna virus probe under low stringency conditions; Alizon et al. (op. cit.) detected no reactivity under similar relaxed conditions. Nonetheless, molecular structural arguments for assigning the AIDS RV to the lentivirus group of mammalian retroviruses are quite compelling. First, the large (9.7 kb) size of the AIDS RV provirus is quite similar to the 10–10.3 kb proviral DNA described for visna (Molineaux and Clements, Gene *23*, 137–148, 1983). Second, the pbs associated with the visna provirus is also tRNA^{lys} (K. Staskus, E. Retzel, A. Haase, and A. Faras, personal communication). Third, similar to the AIDS RV, visna contains a truncated *gag* gene that encodes a p55 *gag* precursor protein and p30, p16, and p14 processed *gag* proteins (Quérat et al., J. Virol. *52*, 672–679, 1984). Fourth, the sizes of the visna *env* glycoproteins (gp150 and gp135) (Quérat et al., J. Virol., op. cit.) are quite similar to the 160 and 120 kd *env* gene products present in AIDS RV infected cells. Finally, visna and other lentiviruses, like the AIDS RV, are exogenous agents since labeled viral DNA probes fail to hybridize to preparations of uninfected host cell DNA. Coupled with the morphological properties and biological characteristics of viral infections described above, the AIDS RVs would seem to be best classified within the lentivirus group of retroviruses.

The Relationship of Different AIDS RV Isolates to One Another

A very important issue in understanding the epidemiology and pathogenesis of AIDS, as well as in developing therapies and vaccine strategies, is the question of the heterogeneity of the virus. How variable is one isolate from another? Where do the variations occur and how will they affect viral pathogenicity and antigenicity?

Two recent reports (Shaw et al., Science *226*, 1165–1171, 1984; Luciw et al., Nature *312*, 760–763, 1984) indicate that AIDS RVs isolated from different individuals exhibit striking structural heterogeneity as monitored by restriction enzyme polymorphisms. In one instance (Shaw et al., op. cit.), 19 of 31 cleavage sites differed. Although the restriction maps of HTLV-III, ARV, and LAV proviral DNAs have not been formally compared, superficial inspection of published cleavage maps and Southern blots suggests that HTLV-III and LAV are closely related to one another whereas ARV and many other isolates are substantially different.

An analysis of AIDS RV nucleotide sequence heterogeneity is presented in Table 1. As a baseline for comparisons, sequence variations between two HTLV-III clones (BH10 and BH8), reported by Ratner et al. (op. cit.), were determined. Small differences ranging from 1.4%–2.2% could be demonstrated in the U3, *env*, and "B" segments. A comparison of HTLV-III and LAV proviruses generated virtually identical results, indicating that HTLV-III was no more different from LAV than molecular clones of HTLV-III were from one another. In contrast, striking differences

were apparent when LAV was compared to ARV. These alterations were primarily concentrated in the 3' half of the AIDS RV. The *env* gene in particular had undergone multiple changes, resulting in a 9.3% nucleotide sequence difference. Further analysis of the LAV and ARV *env* genes indicates: 1) a majority of the nucleotide alterations are located in the 5' half of the *env* coding sequences (and result in nearly 15% nucleotide sequence heterogeneity within this 1300 bp segment), and 2) the differences are primarily a series of reciprocal insertions and deletions (up to 24 bp) in the 5' half of the *env* region of the two proviral DNAs. As shown in Table 1, the alterations in *env* nucleotide sequences in LAV relative to ARV are not trivial. They result in a 21% difference in the deduced amino acid sequence of the two *env* glycoproteins. It is interesting to note that of the 31 and 32 potential glycosylation sites present in the LAV and ARV *env* region, respectively, only 19 are identical.

Although it is presently unclear why infection of man with the AIDS RV results in a slow progressive immunosuppressive disease, some models of lentivirus persistence are consistent with the structural differences observed between LAV/HTLV-III and ARV. Neutralization studies indicate that the periodic nature of disease caused by equine infectious anemia virus (EIAV), a lentivirus, is due to the sequential appearance, in an infected animal, of novel antigenic viral variants that temporarily escape host immune surveillance. The different antigenic strains of EIAV responsible for sequential febrile episodes contain alterations confined to virion glycoproteins, as monitored by tryptic peptide mapping analyses (Montelaro et al., JBC *259*, 10539–10544, 1984). Similar antigenic variants have been reported for visna virus, with changes mapping to the *env* gene (Scott et al., Cell *18*, 321–327, 1979; Clements et al., PNAS *77*, 4454–4458, 1980). However, the relationship of *env* glycoprotein antigenic variations and viral pathogenesis is unclear. In vivo, viral persistence may be due to a restriction of visna gene expression (Haase et al., Science *195*, 175–177, 1977).

The analysis of nucleotide sequence heterogeneity presented in Table 1 indicates that HTLV-III and LAV are virtually identical. This result is surprising in view of their independent isolation and published reports, cited above, which show that extensive restriction enzyme polymorphisms exist among different AIDS RV isolates. To evaluate the significance of the differences in the *env* genes of ARV vs. HTLV-III/LAV, another independent isolate must be sequenced. If this verifies the substantial variability in *env* glycoproteins (reflecting genetic alterations attending AIDS RV infection in man), then preventive and therapeutic strategies may have to be directed to less variable regions of the viral genome.